Accepted for presentation at the IEEE ICDM 2006 Workshop on Data Mining in Bioinformatics (DMB 2006)
draft material -- please do not cite

Predictive Integration of Gene Ontology-Driven Similarity and Functional Interactions

Francisco Azuaje[1,*], Haiying Wang[1], Huiru Zheng[1],Olivier Bodenreider[2] and Alban Chesneau[3]
[1]School of Computing and Mathematics, University of Ulster, UK., [2]National Library of Medicine, National Institutes of Health., USA, [3] High-throughput protein technologies Group, EMBL-Grenoble, France.
E-mail: fj.azuaje@ulster.ac.uk, hy.wang@ulster.ac.uk, h.zheng@ulster.ac.uk, olivier@nlm.nih.gov, chesneau@embl-grenoble.fr

Abstract

The inference of functional networks of genes relies on the integration of multiple data sources and knowledge bases. More generally, there is a need to develop methods to automatically incorporate prior knowledge to support the prediction and validation of novel functional associations. One such important knowledge source is represented by the Gene Ontology (GO)™ and the many model organism databases of gene products annotated to the GO. We investigated quantitative relationships between the GO-driven similarity of genes and their functional interactions by analyzing different types of associations in Saccharomyces cerevisiae and Caenorhabditis elegans. This study demonstrates that interacting genes (including regulatory and protein-protein interactions) exhibit significantly higher levels of GO-driven similarity (GOS) in comparison to random pairs of genes used as a surrogate for negative interactions. Significant associations were identified using annotations from the three GO hierarchies and different GOS techniques. Our study indicates that, the Biological Process hierarchy provides more reliable results for co-regulatory and protein-protein interactions. Statistical analyses suggest that GOS represent a relevant resource to support prediction of functional networks in combination with other resources.

1. Introduction

The reliable prediction of functional networks of genes may be achieved by integrating multiple data sources, such as gene expression, phylogenetic profiles and high-throughput protein-protein interaction experiments. This is necessary because such individual sources may be considered as *weak prediction models* due to their limitations in terms of predictive accuracy and coverage. Several studies have reported significant links between different types of genomic data sets, as well as techniques (e.g. machine learning) to combine them and improve prediction quality for relatively simple model organisms, such as yeast [1], [2]. Furthermore, it is crucial to integrate prior knowledge resources, such as annotation databases and the literature, not only for building advanced functional classifiers, but also to assist in the validation of technique-independent predictions (e.g., to detect potential spurious associations).

The *Gene Ontology ™* (GO) is one such source of prior knowledge, which has become the *de facto* standard for annotating gene products [3]. It has been proposed as a gold standard to assess the quality of several classification systems using, for example, expression data. A comprehensive review of some of such applications has been provided by Khatri and Drăghici [4]. Moreover, information extracted from model organism databases annotated to the GO has been applied for making *de novo* predictions of gene function in relatively simple organisms [5].

Methods based on the GO have been proposed for measuring similarity between genes. Previous research showed significant relationships between GO-driven similarity of pairs of genes and their sequence-based similarity [6]. We have also evaluated relevant quantitative relationships between GO-driven similarity and gene expression correlation [7]. GO-driven clustering algorithms based on such approaches have been recently reported [8]. Moreover, GO-driven similarity provides the basis for developing tools that may facilitate the identification of relevant partitions

from clustering, using, for example, GO-driven cluster validity indices [9].

Prior to integrating a predictive resource, *Res*, it is first necessary to assess its predictive relevance and reliability in relation to data sets of known positive and negative interactions [10]. In this case the hypothesis to prove is: Can information extracted from *Res* be in principle applied to distinguish pairs of interacting genes (positives) from those that have not shown evidence to be interacting (negatives)? Are there significant quantitative relationships to indicate that *Res* may be used as an input to different prediction models?

The application of information from model organism databases annotated to the GO to support the prediction of functional networks of genes has not been rigorously investigated. Jansen *et al.* [1] integrated different genomic data sets including annotations derived only from the GO *Biological Process* hierarchy to predict protein-protein (PP) interactions. The GO-driven similarity of a pair of genes was used as an indicator of PP interactions in yeast. Between-gene similarity was calculated by identifying the set of GO terms shared by the two sets of annotations. For a given database of protein pairs, the total number of protein pairs sharing the same set of annotations was used as an estimator of similarity. Thus, the lower this frequency, the more similar the gene pair under consideration. These authors found that lower term frequencies were correlated with a higher likelihood of finding two proteins in the same complex. Nevertheless, such a similarity assessment approach does not fully exploit relevant topological and information content features that may be useful for meaningfully estimating between-gene similarity. In some cases genes annotated to closely related but distinct GO terms may actually exhibit no similarity according to this method.

Using annotations from the three GO hierarchies: *Molecular Function* (MF), *Biological Process* (BP) and *Cellular Component* (CC), we sought to assess relationships between the GO-driven similarity of a pair of genes and their functional interactions. This study aimed to investigate the feasibility of applying GO-driven similarity to support the prediction of functional interactions of genes, including physical and regulatory interactions, in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. Two key questions addressed were: a) Can GO-driven similarity be applied to estimate the functional coupling of genes, such as gene expression co-regulations and other physical and non-physical interactions and b) Can such a knowledge be used in combination with other resources to improve the prediction process? Our hypothesis is that the GO-driven similarity among genes is a relevant indicator of functional interaction.

The following section describes the data sets analyzed: 1) A data set of annotated co-regulatory interactions from yeast, 2) an extensive, high-quality functional gene network for yeast, and a 3) high-quality PP interaction data set from *C. elegans*. This is followed by a description of the methods applied to measure similarity using GO annotations, its links to the identification of interacting pairs of genes and a statistical assessment of the predictions. Results for the three data sets are presented next. The final section discusses the main contributions of this study, potential applications and future research.

2. Materials

2.1. Data sets

Gene co-regulation in *S. cerevisiae* (CoReg)

This data set originated from a comprehensive collection of annotated regulons compiled by Simonis *et al.* [11]. Their data set comprised more than 1400 pairs of gene-factor associations retrieved from the TRANSFAC [12] and aMAZE [13] databases and literature searches. More than 13000 pairs of co-regulated genes were then extracted from these data. These pairs comprised the CoReg reference data set analyzed in this investigation.

Functional network of yeast genes (FunNet)

This data set was obtained from an extensive, high-quality functional gene network investigated by Lee *et al.* [2]. Unlike the CoReg data set, FunNet comprises different types of functional associations, mediated or not by physical interaction (i.e. protein-protein, regulatory, etc). This network was inferred by integrating diverse, high-quality functional data sets: mRNA coexpression, gene-fusions, phylogenetic profiles, literature co-citation and protein interaction experiments, with the Kyoto Encyclopedia of Genes and Genomes (KEGG) database used as the gold standard. A sub-sample of 19,216 pairs of genes representing the most reliable interaction predictions were analyzed in this study (additional information about FunNet is available in a Supplementary Section that can be obtained on request from the authors).

PP interactions in *C. elegans* (PPInt)

This data set represents another level of complexity, in which 860 protein-protein (PP) interactions, including a few self-interactions, were obtained from the *Worm Interactome* (WI5) map. The selected data

set, from now on referred to as PPInt, contains the highest-confidence, published interactions from WI5 [14].

2.2. The GO

The GO hierarchies – *Molecular Function* (MF), *Biological Process* (BP) and *Cellular Component* (CC) – provide controlled terms for describing the role played by a gene product, the biological goals to which a gene product contributes and the cellular localization of the gene product respectively. Within each hierarchy, GO terms are organized in a *directed acyclic graph*, whose nodes are the terms. There are two types of relationships among GO terms: "is a" and "part of". The first type is used when a child term is more specific than its parent term. The second type is used when a parent has the child as its part. This study takes advantage of both types of links for computing similarity between terms as justified elsewhere [6].

The annotations recorded in the model organism databases consist of associations between gene products and GO terms. The evidence supporting such annotations is captured by evidence codes, including *TAS* (Traceable Author Statement), *ISS* (Inferred from Sequence or structural Similarity) and *IEA* (Inferred from Electronic Annotation). While *TAS* refers to peer-reviewed papers and indicates strong evidence, *IEA* and *ISS* denote automated predictions, i.e., generally less reliable annotations. Further information about databases annotated to the GO and evidence codes for the annotations is available at www.geneontology.org. The reader is also referred to [7] and [15] for an introduction to some of the predictive data analysis applications of the GO.

2.3. GO annotation databases

The pairs of interacting genes in the three data sets presented earlier are annotated to the GO. We performed experiments on data excluding the less reliable annotations (i.e., ignoring annotations whose evidence code is either *ISS* or *IEA*). Moreover, we compared these results against those obtained from excluding only *IEA* annotations. The August 2005 database releases of the *Saccharomyces Genome Database* (SGD) and *WormBase* (WB), all available at www.godatabase.org, provided the GO annotations for these data sets.

CoReg has 8,839, 10,874, and 11,309 interacting pairs with both genes linked to at least one GO term under the MF, BP and CC hierarchies respectively. FunNet had 11,767, 15,520 and 16,865 pairs of interacting genes with both genes associated with at least one GO term under the MF, BP and CC hierarchies respectively. In PPInt the numbers of interacting pairs of genes in which both genes were described by at least one GO term were 152 under the BP hierarchy and 5 under the CC hierarchy. This data set did not contain any valid annotations under MF. The number of annotations reported above refers to non-*ISS*/non-*IEA* annotations. Information about non-*IEA* annotations and a more detailed description of the data sets analyzed is available in a Supplementary Section that can be obtained on request from the authors.

3. Methods

3.1. GO-driven similarity

To estimate the **similarity between two genes** g_k and g_p, annotated with sets of GO terms A_k and A_p respectively, one must first understand how to calculate the similarity between two GO terms. Several i*nformation-theoretic approaches* to measuring ontology-driven similarity have been studied previously [6], [15]. Unlike traditional *edge-counting techniques*, these methods are based on the assumption that the more information two terms share in common, the more similar they are. *Lin's similarity model*, for example, has shown to produce both biologically meaningful and consistent similarity predictions [6], [7] in comparison to related approaches. Given terms $c_i \in A_k$ and $c_j \in A_p$, the between-term Lin's similarity is defined as:

$$sim(c_i, c_j) = \frac{2 \times \max_{c \in S(c_i, c_j)} [\log(p(c))]}{\log(p(c_i)) + \log(p(c_j))} \quad (1)$$

where $S(c_i,c_j)$ represents the set of ancestor terms shared by both c_i and c_j, 'max' represents the maximum operator, and $p(c)$ is the probability of finding c or any of its descendants in the database analyzed. It generates normalized values between 0 and 1.

Between-gene similarity results from the aggregation of similarity values between the annotation terms of these genes. In practice, given a pair of gene products, g_k and g_p, with sets of annotations A_k and A_p comprising m and n terms respectively, the between-gene similarity, $SIM(g_k, g_p)$, is defined as the *simple average* (inter-set) *similarity* between terms from A_i and A_j:

$$SIM(g_k, g_p) = \frac{1}{m \times n} \times \sum_{c_i \in A_k, c_j \in A_p} sim(c_i, c_j) \qquad (2)$$

where $sim(c_i,c_j)$ may be calculated using (1). Nevertheless, this method might not always produce consistent results. For example, intuitively, the similarity between two genes having the same sets of annotation terms is expected to be equal to 1. However, this is not true when several annotations within a hierarchy are assigned to the genes. It will estimate, for instance, $SIM(g_i, g_j) = 0.5$, for $g_i = g_j$ when A_i and A_j are described by the same set of annotations with more than one GO term within a hierarchy. In order to address this limitation, we have introduced an alternative approach that selectively aggregates *highest average* (inter-set) *similarity* values [15] as follows:

$$SIM(g_i, g_j) = \frac{1}{m+n} \times \left(\sum_k \max_p (sim(c_k, c_p)) + \sum_p \max_k (sim(c_k, c_p)) \right) \qquad (3)$$

These approaches and their relationships to sequence-based similarity and co-expression have been investigated in [6] and [7]. From now on we will refer to (2) and (3) as the *simple* and *highest average similarity* methods respectively.

3.2. Linking GO-driven similarity and functional interactions

Comparing GO-driven similarity to other indicators of functional relations. GO-driven similarity values were calculated for all the annotated pairs of genes in the data sets described in Section 2. These data represented our sets of true positive interactions, which were statistically analyzed to show significant relationships with GO-driven similarity. In order to illustrate such links, similarity values from these sets of true positive interactions were compared to similarity values measured in a set of randomly associated genes, used as a surrogate for negative interactions, i.e. pairs of genes not showing evidence of interaction. In practice, a set of "non-interacting genes" was produced as follows. For a given data set, **P**, comprising *M* true positive interactions, a set **N**, with *M* negative interactions was built by randomly pairing genes from **P**. Moreover, the resulting sets were verified to ensure that newly formed pairs were not included in **P**. One has to take into account that some of the pairs included in **N** may actually be false negatives (i.e., interacting genes whose interaction has not been not recorded in **P**) and this might influence the comparisons performed. However, at least with regard to the data sets analyzed (evidence available) this could not be demonstrated. The resulting data sets **N** represent a valid approximation of counter-examples, which are essential to explore potential associations between functional interactions and GO-driven similarity. Furthermore, the random effects and variability linked to this data sampling procedure is reduced by generating *K* independent **N** sets. These *K* sets are then analyzed as an aggregated set, **N'**, consisting of *K* x *M* pairs of (non-interacting) genes.

Fundamental relationships between GO-driven similarity and the existence/absence of interactions were estimated by comparing similarity values exhibited in **P** versus values observed in **N'**. This was done for each of the three problems described in Section 2 and for the three GO hierarchies independently. Differences between **P** and **N'** were summarized by estimating their respective mean similarity values. The significance of their differences was tested by applying the Student's *t*-Test. The relevant null hypothesis tested was that these mean similarity values originated from the same sample, i.e. there are no significant differences between mean values in **P** and **N'**. This relatively simple comparison provided key insights into relationships between the degree of similarity of pairs of genes and the likelihood that these genes are functionally interacting.

Using GO-driven similarity to predict interactions. After identifying significant differences, the capacity of GO-driven similarity to predict functional interactions (as a single predictive source) was analyzed. Given a similarity value, $SIM(g_k, g_p)$, and a pre-defined *predictive similarity threshold* value, *GOS-Th*, genes g_k and g_p are predicted to be an interacting pair (positive interaction) if $SIM(g_k,g_p) \geq GOS\text{-}Th$. Some of these predictions will obviously be false. Therefore, the next task was to estimate the rate of falsely predicted interactions. More generally, this is related to the problem of estimating the *decisive false discovery rate*, which has shown to be a robust and conservative estimator of the probability, *P*, of detecting spurious associations [16]. To estimate *P*, *AbN'* and *AbP* are calculated. *AbN'* represents the number of interactions that would occur by chance and *AbP* the number of pairs correctly predicted as positive interacting pairs. The ratio *AbN'*/*AbP* represents the rate of falsely predicted interactions. *AbN'* was estimated using the mean number of interacting pairs obtained from the *K* data sets, **N**, i.e. the total number of interactions observed in **N'**, divided by *K*. A rate of falsely predicted interactions, *P*, close to 1 corresponds to random prediction. In contrast, low *P* values indicate strong evidence to support the validity of the

positive interactions detected by the GO-driven similarity method. *P* values were calculated for the data sets described above using different *GOS-Th* values. This analysis allows one to have a better idea about how many false positive predictions may potentially be made when applying the GO-driven similarity method as a single prediction model. The analysis tasks described above were carried out with *K* = 10. Limited computing power was the reason for not generating larger numbers of sets **N**. However, the relatively large number of gene pairs included at least in the first two data sets should contribute to the reduction of the bias and variability of the estimations

4. Results

4.1. Results from CoReg

Table 1 summarizes the differences between the sets **P** (positives) and **N'** (negatives) with regard to their mean similarity values from the simple and highest average methods respectively. Unknown, *IEA* and *ISS* annotations were excluded. Interacting pairs of genes generally exhibit higher similarity values than non-interacting pairs using both methods. The high *t* values obtained suggest significant differences ($p < 0.001$) for all GO hierarchies. This suggests the feasibility of applying GO-driven similarity to support the distinction of co-regulated from non-co-regulated pairs of genes. Figure 1 shows the estimated probabilities, *P*, that such predictions are false as a function of the predictive threshold, *GOS-Th*.

Table 1 CoReg: Differences between interacting and random, non-interacting pairs of genes in terms of their GO-driven similarity. SE: standard error of the estimated mean.

Simple average similarity method			
GO Hierarchy	**True Positives (Mean ± SE)**	**Random Pairs (Mean ± SE)**	**Significance**
MF	0.16 ± 0.002	0.10 ± 0.0005	$p < 0.001$
BP	0.27 ± 0.003	0.10 ± 0.0005	$p < 0.001$
CC	0.34 ± 0.003	0.20 ± 0.0006	$p < 0.001$

Highest average similarity method			
GO Hierarchy	**True Positives (Mean ± SE)**	**Random Pairs (Mean ± SE)**	**Significance**
MF	0.16 ± 0.002	0.10 ± 0.0005	$p < 0.001$
BP	0.27 ± 0.003	0.13 ± 0.0005	$p < 0.001$
CC	0.34 ± 0.003	0.25 ± 0.0008	$p < 0.001$

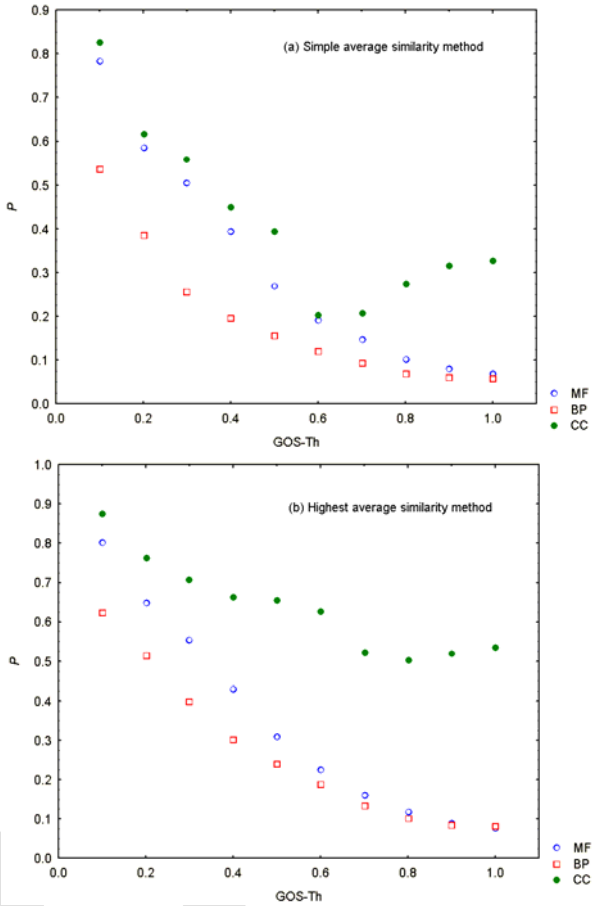

Figure 1 CoReg: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

4.2. Results from FunNet

Tables 2 summarizes the differences between the sets **P** and **N'** in terms of their mean similarity values using the simple and highest average methods respectively. The high *t* values obtained suggest significant differences ($p < 0.001$) for all GO hierarchies. Unknown, *IEA* and *ISS* annotations were excluded. With both methods, pairs of interacting genes tend to exhibit higher similarity values than pairs of non-interacting genes. This suggests the feasibility of using GO-driven similarity to help to distinguish interacting from non-interacting genes (including physical and non-physical interactions). Figure 2 shows the estimated probabilities, *P*, that such predictions are false as a function of *GOS-Th*.

Table 2 FunNet: Differences between interacting and random, non-interacting pairs of genes in terms of their GO-driven similarity. SE: standard error of the estimated mean.

Simple average similarity method

GO Hierarchy	True Positives (Mean ± SE)	Random Pairs (Mean ± SE)	Significance
MF	0.62 ± 0.004	0.26 ± 0.0010	$p < 0.001$
BP	0.58 ± 0.003	0.28 ± 0.0008	$p < 0.001$
CC	0.58 ± 0.002	0.34 ± 0.0007	$p < 0.001$

Highest average similarity method

GO Hierarchy	True Positives (Mean ± SE)	Random Pairs (Mean ± SE)	Significance
MF	0.66 ± 0.004	0.27 ± 0.0010	$p < 0.001$
BP	0.65 ± 0.003	0.31 ± 0.0008	$p < 0.001$
CC	0.64 ± 0.002	0.37 ± 0.0007	$p < 0.001$

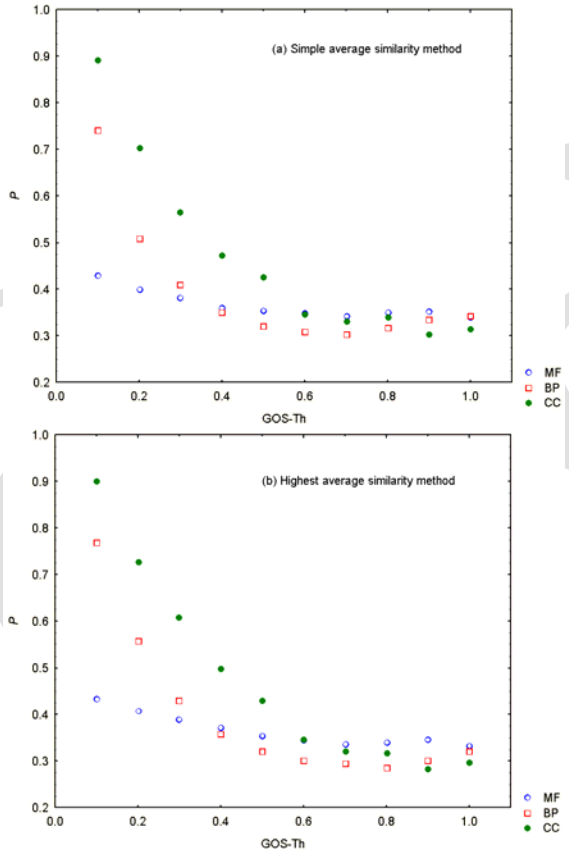

Figure 2 FunNet: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

4.3.Results from PPInt

Tables 3 summarizes the differences between the sets **P** and **N'** with regard to their mean similarity values using the simple and highest average similarity method. Unknown, *ISS* and *IEA* annotations were excluded from this analysis. Significant differences ($p < 0.05$) were observed only in connection to the BP hierarchy. Figure 3 presents the estimated probabilities, *P*, that such predictions are false as a function of *GOS-Th*. The interpretation of these figures should also take into account the very low number of gene pairs with CC annotations.

Table 3. PP-Int: Differences between interacting and random, non-interacting pairs of genes in terms of their GO-driven similarity. SE: standard error of the estimated mean. N.S: no significance

Simple average similarity method

GO Hierarchy	True Positives (Mean ± SE)	Random Pairs (Mean ± SE)	Significance
MF	NA	0.28 ± 0.200	NA
BP	0.23 ± 0.02	0.19 ± 0.005	$p < 0.05$
CC	0.37 ± 0.19	0.22 ±0.040	N.S

Highest average similarity method

GO Hierarchy	True Positives (Mean ± SE)	Random Pairs (Mean ± SE)	Significance
MF	NA	0.30 ± 0.200	NA
BP	0.45 ± 0.02	0.39 ± 0.007	$p < 0.05$
CC	0.38 ± 0.19	0.24 ± 0.040	N.S

5. Discussion and Conclusions

Relationships between GO-driven similarity, based on an information theoretic approach, and different levels of functional interaction were investigated. Three complexity levels were explored: Co-regulation in *S. cerevisiae*, a more comprehensive set of functional associations (including both physical and non-physical interactions) in the same organism and a smaller set of PP interactions in *C. elegans*.

Data set and method selection. We focused our investigation on previously published, high-quality annotated interactions, which represented the reference data sets for this study. The less reliable annotations were removed from the data used to compute GO-driven similarity. We concentrated on a GO-driven similarity assessment approach (Lin) that has previously shown to be strongly related to sequence-based similarity and gene co-expression [6], [7].

Moreover, we assessed the highest average similarity method that aims to improve estimation consistency.

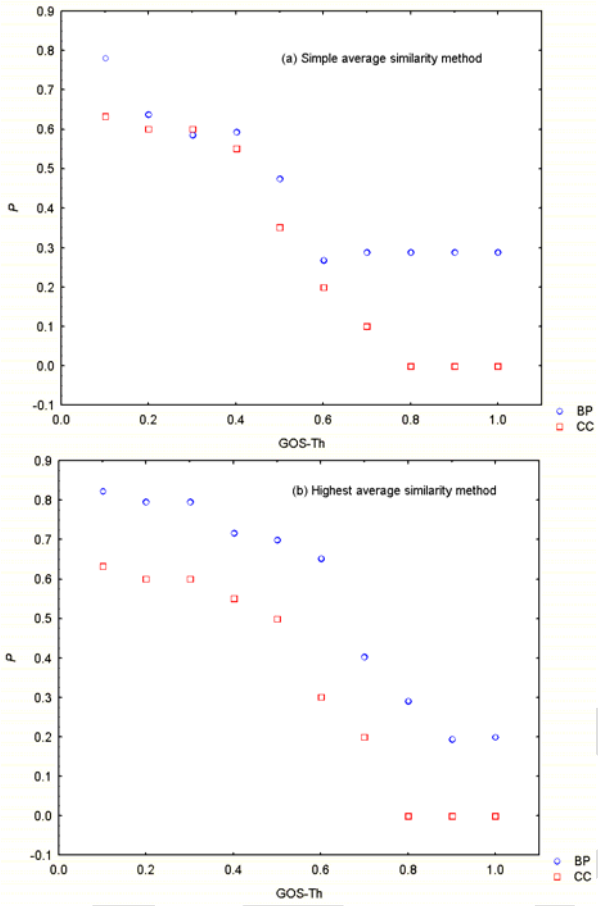

Figure 3. PP-Int: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

Summary of findings. This study demonstrated significant relationships between functional similarity (assessed by GO-driven similarity) and known interactions. This pattern was remarkably observed under all hierarchies for CoReg and FunNet. This supports the hypothesis that the GO-driven similarity of pair of genes may be applied to support the prediction of functional interactions (including co-regulatory and PP interactions) in yeast. Furthermore, by only excluding IEA annotations (including ISS) similar general trends can be observed in all data sets, similarity estimation methods and GO hierarchies. This comparison is included in a Supplementary Section along with additional experiments suggesting that the degree of GO-driven similarity is consistent with interaction likelihood scores of pairs of genes as reported by Lee *et al.* [2] based on a comprehensive, integrative prediction strategy. The Supplementary Section can be obtained on request from the authors. We also performed a manual verification (details in Supplementary Section) to assess the potential biological significance of some of the "false positive" (novel) links. This procedure reported nine pairs of proteins (with unknown interactions), which are feasible candidates to be interacting partners in *C. elegans*, such as F28D1.2 and B0547.1, which are involved in DNA repair and ubiquitination, respectively.

Applications to interactions prediction. Our research does not of course suggest that this approach is sufficient or even necessary to detect relevant interactions. Similarly, we did not aim to argue that it represents a more effective prediction model than existing approaches. However, this investigation offered evidence to motivate the application of this functional similarity measure as a *complementary predictive resource* of functional interaction. This, in combination with other sources, such as gene co-expression and different interaction prediction models, may support more accurate and biologically-meaningful predictions. Integrative prediction models such as those reported by Lee *et al.* [2] and Jansen *et al.* [1] may benefit from incorporating this knowledge-based source. The predictive performance of emerging models should be thoroughly assessed using, for example, receiver-operator characteristic (ROC) curves and other powerful statistical tests, which is also part of our current and future research (see below).

Applications to annotation prediction. The results highlight two important protein groups. The first group represents protein pairs that are not similar based on their GO terms. These proteins may be true negatives, but they may also represent interacting proteins for which no interaction has been recorded yet. This could be the case for proteins with incomplete GO annotations or for those involved in processes not properly described by this ontology. The second group corresponds to similar proteins pairs in relation to their GO-driven annotations. As demonstrated above, such similarity offers strong evidence for the presence of functional interaction. Functional similarity can also be combined with other post-genomic sources of evidence, such as genome-wide in-situ hybridization, to improve the detection of false positive interactions. Thus, the GO-driven similarity approach can also complement experimental approaches to determining false positives. In the case of metazoan organisms, e.g. *C. elegans*, the GO-driven similarity approach is much more difficult to assess as the function of a protein can be related to its tissue- or organ-specific expression patterns. It would be important to integrate gene

expression and tissue localization information to complement GO-driven similarity. Moreover, it might be possible to infer tissue localization from GO annotations. This dimension, which is not considered in unicellular organisms, underlies the complexity of this prediction task and the importance of implementing integrative, module-based approaches to interactome prediction.

Related work. P.H Lee and D. Lee [17] recently integrated ontology-driven similarity information as part of their *modularized network learning method* (MONET). They first identified modules of interrelated genes using gene expression correlation and MIPS (Munich Information center for Protein Sequences database) annotations. *Bayesian networks* were then inferred from the detected modules that successfully predicted relevant gene regulation networks in yeast. Ontology-driven similarity was used to aid in the identification of clusters of genes on the basis of their MIPS annotations. Between-gene similarity was estimated using the between-term Resnik's method [18], which is also an information-theoretic approach. But unlike Lin's, Resnik's method generates un-normalized similarity values ranging from 0 to infinity. Moreover, previous research has shown that Lin's technique may outperform Renisk's and other information-theoretic approaches [19]. For example, Lin's method generates similarity values highly correlated with human assessments of similarity in different application domains. It also reflects significant relationships with gene co-expression. Such relationships are represented in a more consistent and meaningful fashion in comparison to Resnik's approach [7].

Wang *et al.* [8] proposed a GO-driven hierarchical clustering method based on Lin's technique, which identified significant functional modules relevant to several responses to stimuli in yeast. Their method may complement P.H. Lee and D. Lee's method [17] for the detection of functional modules based on GO annotations. A recent study on global prediction of regulatory networks in yeast found that more than 12% of genetic interactions included genes with identical GO annotations [20]. It also found that over 27% of the interactions comprised similar annotations sets based on a conservative estimate of similarity, which approximated the degree of annotation overlaps. Our investigation provides further evidence of the potential of GO-driven similarity information to facilitate the prediction of functional interactions. Moreover, we have shown that these relationships go beyond the regulatory level and can support applications involving uni- and multi-cellular organisms.

Limitations. The similarity value distributions, significant differences and the potential to reduce the probability of detecting spurious associations encourage further investigations. Moreover, we believe that, even when significant results were obtained, our assessment may actually be under-estimated. This is because many of the pairs of genes included in the randomly-generated sets ("true negatives") might indeed be part of more comprehensive collections of true positive interactions not included in this study. Some of them might also become true positive interactions in the future with the emergence of new experimental and validated evidence. This factor also suggests that the differences and relationships identified could be stronger. The relatively large amount of interacting pairs included in the Co-Reg and FunNet data sets may contribute to the reduction of such noise sources and possible bias.

One of the major challenges in predicting interaction maps is to remove false-positives interactions. False positives predictions can be caused by technical or biological factors. The former have no biological meaning and come from technical limitations such as the number and the strength of phenotypic tests used for two-hybrid screenings or purification steps realized for complex identification by mass spectrometry. The latter may originate from proteins that actually interact but are not expressed in the same tissues or organs. For example, it was estimated that about 25% to 50% of the genome-wide PP interaction predictions reported in many high-profile publications actually represent false positive interactions [21]. This investigation also estimated probability values, P, that offered relevant insights into these relationships. These indicators help us to further assess the predictive ability of the GO-driven similarity method to detect valid interactions for different prediction similarity thresholds, *GOS-Th*. The results suggest that in general the larger the *GOS-Th*, the lower the probability of making false positive predictions. But it also highlights the fact that many of the false positive interaction predictions might show relatively high similarities. This may also be explained by the difficulties in creating exact true negative data sets. Nevertheless, the results strongly suggest that there is a tendency to reduce the number of false positive interactions by applying more rigorous thresholds. With regard to PPInt, any interpretation of Figure 3 should take into account the very low number of gene pairs with CC annotations. In this case the results only suggest the significance of GO-driven predictions based on the BP hierarchy. The results also confirm that the higher the *GOS-Th*, the more limited the predictive coverage of the model, i.e. the higher the

possibility of missing true positive interactions. This property is perhaps more visible in the CoReg and FunNet data sets. Alternative assessments may incorporate other estimators of *P*, including less conservative methods. Tables 1 to 3 and Figures. 1 to 3 confirm that, in principle, it would be possible to implement more accurate predictive models with higher *GOS-Th* values, but at the risk of reducing predictive coverage. Predictions based on annotations from the BP hierarchy are in general indicated as reliable and accurate for the three data sets analyzed. Similar results were obtained when including ISS annotations. With regard to the CC hierarchy, the relatively high *P* values and their non-monotonic response against *GOS-Th* confirms its lack of reliability as a predictor of positive interactions for CoReg. These results might not be considered as surprising findings. However, they highlight the potential of using GO-driven similarity as an alternative *weak prediction model*, which may complement other weak predictive resources, e.g. gene co-expression and high-throughput interaction identification techniques. It also opens opportunities to incorporate prior knowledge to support the automated assessment of predictions derived from large-scale studies.

Future work. We are currently applying the GO-driven similarity assessment approach to support the prediction of integrated, large scale functional networks in different model organisms [22]. We will also compare our methods with recent contributions in this area [23].

Acnowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). This paper was revised while FA was a visiting scholar at the Lister Hill National Center for Biomedical Communications, NLM, NIH.

REFERENCES

[1] R. Jansen, H. Yu, D. Greenbaum, and *et al.* "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol.302, 2003, pp.449-453.

[2] I. Lee, S. V. Date, A. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol.306, 2004, pp.1555-1558.

[3] The Gene Ontology Consortium, "Creating the gene ontology resource: Design and implementation," *Genome Research*," vol.11, 2001, pp.1425-1433.

[4] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol.21, 2005, pp.3587-3595.

[5] O. D. King, R. E. Foulger, S. S. Dwight *et al.* "Predicting gene function from patterns of annotation," *Genome Research*, vol.13, 2003, pp.896--904.

[6] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, 2003, pp.1275--1283.

[7] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, "Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships," In *Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, La Jolla, CA, USA, 2004, pp.25-31.

[8] H. Wang, F. Azuaje, and O. Bodenreider, "An ontology-driven clustering method for supporting gene expression analysis," In *Proc. of the 18th IEEE International Symposium on Computer-Based Medical Systems,* 2005, pp.389-394.

[9] N. Bolshakova, F. Azuaje**,** and P. Cunningham, "A knowledge-driven approach to cluster validity assessment," *Bioinformatics*, vol.21, 2005, pp.2546-7.

[10] R. Jansen and M. Gerstein, "Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction," *Curr Opin Microbiol*, vol.7, 2004, pp.535-45.

[11] N. Simonis, S. J. Wodak, G.N. Cohen, and J. van Helden, "Combining pattern discovery and discriminant analysis to predict gene co-regulation," *Bioinformatics*, vol.15, 2004, pp.2370-2379.

[12] E. Wingender, . X. Chen, R. Hehl, and *et al.* "TRANSFAC: an integrated system for gene expression regulation," *Nucleic Acid Res*., vol.28, 2000, pp.316-319.

[13] J. van Helden, A. F. Rios, and J. Collado-Vides, "Discovering regulatory elements in non-coding sequences by analysis of spaced dyad," *Nucleic Acid Res*., vol.28, 2000, pp.1808-1818.

[14] S. Li, C. M. Armstrong, N. Bertin, and *et al.* "A map of the interactome network of the metazoan *C. elegans*," *Science*, vol.303, 2004, pp.540-543.

[15] F. Azuaje, H. Wang, and O. Bodenreider, "Ontology-driven similarity approaches to supporting gene functional assessment," In *Proc. Of The Eighth Annual Bio-Ontologies Meeting*, Michigan, 25 June, http://bio-ontologies.man.ac.uk/.

[16] D. R. Bickel, "Probabilities of spurious connections in gene networks: application to expression time series," *Bioinformatics*, vol. 21, 2005, pp.1121-1128.

[17] P. H. Lee and D. Lee, "Modularized learning of genetic interaction networks from biological annotations and mRNA expression data," *Bioinformatics*, vol.21, 2005, pp. 2739-2747.

[18] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," In *Proc. of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 1995, pp.448-453.

[19] D. Lin, "An information-theoretic definition of similarity," in *Proc. of 15th International Conference on Machine Learning*, San Francisco, 1998, pp.296-304.
[20] A. H. Tong, and *et al.* "Global mapping of the yeast genetic interaction network," *Science*, vol.303, 2004, pp.808-813.
[21] A. M. Edwards, B. Kus, R. Jansen, and *et al.* "Bridging structural biology and genomics: assessing protein interaction data with known complexes," *Trends in Genet*ics, vol.10, 2002, pp.529-536.
[22] F. Browne, H. Wang, H. Zheng, F. Azuaje, "An assessment of machine and statistical learning approaches to inferring networks of protein-protein interactions", *Journal of Integrative Bioinformatics,* 2006, in press.
[23] X. Wu, L. Zhu, J. Guo, D.Y Zhang, and K. Lin, "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations", *Nucleic Acids Res.*, vol. 34, 2006, pp. 2137-50.